Semantic integration to characterize microbial pathogens: Multi-resource enrichment of experimental proteomic and genomic datasets

Erich Gombocz ¹), James Candlin²)

Summary

Bacterial and viral caused infectious diseases account for major health threats globally, yet the characterization, identification and understanding of them has scientifically challenged. This is mainly due to the fact that while there is a wealth of information (and even complete genomes) available, its integrated utilization in context of the biological system to better understand causes and similarities in infectious diseases is still in its infancy. This work tries to address some of the many obstacles involved in this endeavor as it attempts to identify peptides from different microorganisms with common mechanisms of actions causing disease, and to use them as biomarkers to detect pathogenic microbial threats prior to onset of disease symptoms to help in outbreak prevention.

The workflows to accomplish this goal consist of 5 steps. The first step is a thorough peptide analysis of microorganism via mass spectrometry and their identification by sequence scoring (Sorcerer, indexed SEQUEST search, BioWorks). The second step is the annotation of peptides with genes and genomic sequences relevant to protein expression to qualify the accuracy of the identification. Step 3 involves the use of public domain microbial databases (PATRIC, MIST, ICTV, VIDA, Viral ORFeome, miRBase) to semantically integrate the experiments with organism taxon-specific functional domains and pathway information relevant to diseases caused by the pathogens. Based on sequence similarity, sequences are clustered into homologous protein families (HPFs), and those protein families are enriched with annotations including functional classification, related protein structures, homologous protein length, boundaries of conserved regions and bacterial or virus homologous protein identifications. Further enrichment is achieved through addition of disease-related pathways (BioCyc, KEGG). The resulting knowledgebase provides a network with functional annotations to peptides and their relationships to diseases (Sentient Knowledge Explorer). In Step 4, those peptides in the network are identified which have similar disease-causing functions and appear in several pathways. Intergrogating the network via semantic queries (SPARQL) results in discovery of key pathway intersections commonly involved in the disease. The last step is the creation of molecular marker signatures (SPARQL, Applied Semantic Knowledgebases - ASK) and test their validity as disease markers.

Future applications will apply this technology for rapid detection of biological threats, to characterize origin and type of disease outbreaks and to develop preventive measures (such as broadly applicable drugs or vaccines) effective for entire classes of pathogenic organisms.

Materials and Methods

• Typsin digests of abundant protein depleted Human serum were separated via Gradient LC-MS/MS (LTQ Orbitrap Velos, Thermo Fisher). MS/MS peptide spectra were analyzed to identify microbes and score matches against a database of pathogenic microbial sequences (ABCD traceback). Automated spectrum processing was performed via a set of pre-configured tools (Sorcerer, SEQUEST, Trans-Proteomic Pipeline (TPP) on the Sorcerer™ analysis platform (SAGE-N)).
• A pre-loaded microbial knowledgebase (IO Informatics) enriched with harmonized public resources (PATRIC, MIST, ICTV, VIDA, Viral ORFeome, miRBase, BioCyc, KEGG) under a common ontology was utilized for semantic enrichment on the experimental peptide lists (Sentient Knowledge Explorer™).
• Datasets were enriched with functional peptide and gene annotations in a systems biology network for microbial pathogens and analyzed for marker pattern with common disease pathway relationships to be used for decision support in screening.

Results

• Using a pre-configured analytical pipeline to identify peptides from LC-MS/MS of serum samples containing microbial pathogens, it was possible to identify microbial peptide sequences and distinguish them from host peptides at ~95% sequence scoring (Step 1, Sorcerer™ Platform). Exemplary peptides obtained in this way are listed in Table 1.
• Independently, a semantic microbial knowledgebase was created using a broad list of known bacterial and viral peptides and output from public microbial databases (PATRIC, MIST, ICTV, VIDA, Viral ORFeome, miRBase) under a common, dynamically established application ontology. Harmonization was achieved through use of thesauri for microorganisms and diseases during the semantic import mapping and merging. This knowledgebase was further enriched with disease-related pathway information relevant to the pathogen peptides and their genomic annotations (BioCyc, KEGG), providing a robust semantic and systems biology network of microbial pathogens (Step 2, Sentient Knowledge Explorer™, OpenLink Virtuoso Universal Server).
• Similarly, the peptide list from Table 1 was imported and the peptides visualized and analyzed for intersections between disease pathways of pathogenic organisms (Fig.2). Visual SPARQL queries identify peptides with similar disease-causing functions appearing in several pathogens. Results are discovery of key pathway intersections commonly involved in the disease. Iterative refinement of peptide pattern leads to molecular marker signatures (SPARQL, Applied Semantic Knowledgebases - ASK™), directly applicable for microbial pathogen screening.

Discussion

• This study focused on establishing initial sets of microbial peptide markers obtained from sequence-matched LC-MS/MS spectra using a pre-configured analytical pipeline and semantic knowledgebase (Sentient Knowledge Explorer). Objective was to check the validity of such an approach to rapidly identify biological marker pattern applicable to pathogen screening. The presented results consent that notion.

It should be noted that applying semantic technology was instrumental to the success. However, while this study represents a very promising step towards marker-based rapid microbial pathogen detection, it also should be emphasized that additional work will be required in experimental validation to assure broad applicability.

Applying semantic technology to the integration of experimental and public knowledge resources provides a rapid, cost-effective, extensible and biologically sound method towards understanding of disease mechanisms of microbial pathogens.

Once fully validated across a large sample sets, the peptide marker signatures of microbial pathogens will lead to the development of low-cost multiplexed assays for rapid detection of biological threats, to characterize origin and type of disease outbreaks and to develop preventive measures (such as broadly applicable drugs or vaccines) effective for entire classes of pathogens.

References

(5) M. Mar Alba, M. L. D. Pear, S. A. P. S. Gygi, E. Gombocz: “Bioinformatics ORFeome, miRBase) to semantically integrate the experiments with organism taxon-specific functional domains and pathway information relevant to diseases caused by the pathogens. Based on sequence similarity, sequences are clustered into homologous protein families (HPFs), and those protein families are enriched with annotations including functional classification, related protein structures, homologous protein length, boundaries of conserved regions and bacterial or virus homologous protein identifications. Further enrichment is achieved through addition of disease-related pathways (BioCyc, KEGG). The resulting knowledgebase provides a network with functional annotations to peptides and their relationships to diseases (Sentient Knowledge Explorer). In Step 4, those peptides in the network are identified which have similar disease-causing functions and appear in several pathways. Intergrogating the network via semantic queries (SPARQL) results in discovery of key pathway intersections commonly involved in the disease. The last step is the creation of molecular marker signatures (SPARQL, Applied Semantic Knowledgebases - ASK) and test their validity as disease markers.

A merged knowledgebase network in Sentient Knowledge Explorer:

• Left: Dynamic application ontology and instances;
• Bloom: Relationship browser;
• Center: Network graph

Yellow icons depict diseases; pathogens are shown in blue; peptide ions or virus; blue gene symbols represent the genes and orange cycles characterize the pathways.

To reveal resource provenance, some of the total of 15 public databases are shown with at least one of their source connections.

* Certain sample, peptide and organism labels are intentionally hidden in this network to protect ongoing research.

Middle: Table 1. Exemplary peptide output from Sorcerer™

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Protein name</th>
<th>Protein id</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>pld_456</td>
<td>Pld protein</td>
<td>AB123456</td>
<td>Virus</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Final payoff:

• Visual model, iterative refinement
• Qualification of peptide markers
• Result patterns directly applicable decision support in screening for biological threats.

Fig 1: Workflow from sample to microbial pathogen signature

Fig 2: Microbial Knowledgebase in Sentient Knowledge Explorer™. Semantically integrated network of 15 different public resources and experimental peptide results

Fig 3: Visual SPARQL queries (from left to right: a. Molecular parameters for pathogens causing similar disease; b. Qualification of a single microbial peptide sequence for applicability as indicator for a group of pathogens; and c. Verification of biological validity of adding the peptide form b to a set of 7 peptide markers as classifier)